DOCUMENT RESUME

ED 185 116 TM 800 140

AUTHOR Serlin, Ronald C.; Levin, Joel R.
TITLE Coding for Direct Interpretation in Multiple Regression Analysis.
PUB DATE Apr 80
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Covariance; *Analysis of Variance; Aptitude Treatment Interaction; Mathematical Formulas; Mathematical Models; *Multiple Regression Analysis; *Predictor Variables; Statistical Significance
IDENTIFIERS Dummy Variables; Regression Analysis

ABSTRACT

A general procedure is presented for generating code values for a qualitative variable in multiple linear regression analyses that result in directly interpretable estimates of interest. The basic approach, in viewing ANOVA as a multiple regression problem, is to derive quantitative code values for the various levels of the qualitative ANOVA factors. The effect of different coding schemes is illustrated by a simple two-group problem. These notions are then extended, with accompanying directions for code value choices, to the K-group ANOVA problem, the analysis of covariance, and the aptitude-by-treatment interaction (ATI) situation. The ANCOVA model differs from the ANOVA model only in that one or more quantitative covariates are added to the regression equation. In ATI studies the estimated parameters may be directly interpreted as differences in slope values between groups by using these methods. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED185116

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Ronald C. Serlin

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Coding for Direct Interpretation in Multiple Regression Analysis

Ronald C. Serlin and Joel R. Levin

University of Wisconsin

Running Head: Multiple Regression Coding

TM800140

Coding for Direct Interpretation in Mutliple-Regression Analysis

Ronald C. Serlin and Joel R. Levin

University of Wisconsin

ABSTRACT

Multiple linear regression is a versatile model for encompassing analysis of variance, analysis of covariance, and aptitude-by-treatment interaction designs. The question of how to code levels of a qualitative variable is addressed in this paper. Although a variety of coding schemes will produce invariant omnibus statistical results for a given set of data, one's interpretation of treatment effects and treatment differences depends on the particular code values chosen. A general procedure is presented that allows the user to generate, on an a priori basis, code values that result in directly interpretable estimates of interest.

Coding for Direct Interpretation in Multiple Regression Analysis

Ronald C. Serlin and Joel R. Levin

University of Wisconsin

In recent years, multiple linear regression has captured the fancy of behavioral researchers as a versatile data analysis technique that is capable of incorporating the traditional analysis-of-variance (ANOVA) models and more. The basic approach, in viewing ANOVA as a multiple regression problem, is to derive quantitative code values for the various levels of the typically qualitative ANOVA factors. Then, once having coded these factors quantitatively, one can determine their relationship to the quantitative dependent variable. Statistical tests of the relationship are synonymous with inferences about treatment differences or differences among levels of the ANOVA factor(s). For very readable accounts of these procedures and their corresponding interpretation, see Cohen & Cohen (1975) or Kerlinger & Pedhazur (1973).

Surprisingly, perhaps, to the statistical novice, any number of coding schemes can be selected to produce identical results, as far as overall tests of significance are concerned. In fact, Cohen (1968) has shown that even "nonsense" coding schemes produce the same overall test statistics. Given that this is the case, then, is there a basis for choosing one particular coding scheme from among those offered in the previously cited texts? As will be shown here, the choice of a coding scheme is equivalent to the choice of a set of planned ANOVA contrasts. Just as contrasts are usually chosen to provide both interpretable estimates and statistical tests of these estimates, we argue that so should multiple regression coding schemes. Although both Cohen and Cohen (1975) and Kerlinger and Pedhazur (1973) emphasize that the

various coding schemes provide tests of certain effects, often the estimation of the magnitude of these effects is a byproduct whose relationship to the original code values is not readily apparent.

In traditional analyses of variance, contrasts of interest may be computed directly from the cell means. Unfortunately, the same cannot be said of contrasts in the multiple regression approach to ANOVA, since the cell means are not always available. Moreover, as the models considered here become more complex, the connection between coding schemes and parameters estimated can become tenuous and even lost. It will be shown that in these more complex designs, the importance of one's code value choices cannot be over-emphasized.

Two-Group Problem

To provide a context for the discussion that follows, and to illustrate differing coding schemes, let us consider a simple two-group problem. To determine whether the mean difference is statistically significant, one could perform a one-way ANOVA, which is equivalent here to a two-sample t test. Alternatively, the question could be rephrased in terms of detecting a significant association between Y, the dependent variable, and X, a quantitatively coded group factor. The resulting correlation (a point biserial r), if tested for significance, would yield precisely the same numerical result as the previously mentioned t test. Since K = 2 groups, one would have to select a different coded X value for each group in computing the point biserial r. But which values of X to select? Any pair of different values will yield the same statistical result but, as indicated above, a researcher would be wise to select that pair of X values which provides direct estimates of the parameters of interest.

It is easy to illustrate these notions. Consider the three following code value choices:

(a) $X_1 = 0$, $X_2 = 1$

(b) $X_1 = -1$, $X_2 = +1$

(c) $X_1 = -.5$, $X_2 = +.5$

and their corresponding graphical representations in Figure 1. Code choice a

---

Insert Figure 1 about here

---

represents what has been termed a dummy code; b is called a contrast code; and c is proportional to a contrast code. Now if Y were predicted from X in each case, a regression analysis would yield an intercept (B_0) and a slope (B_1) for the regression line. As indicated above, even though the statistical test, H_0: $\beta_1 = 0$, would yield the same results in all cases, the values of B_0 and B_1 will differ from one case to the next. This is clearly evident from Figure 1. For a researcher who desires that B_1 represent an estimate of the difference between the two group means and B_0 an estimate of Group 1's mean, code choice a is appropriate. For the same interpretation of B_1 and for B_0 representing the unweighted mean of the two groups combined, code choice c is appropriate. Finally, code choice b provides an estimate of the unweighted overall mean through B_0, but B_1 is only proportional to the mean difference. Thus, if one wished to interpret B_1 directly as the mean difference, code choices a or c would permit this, but code choice b would not. Of course, in this simple example based on only two groups, the transformation of code choice b to the proper interpretation may seem trivial. For $K > 2$, with complex comparisons and/or not necessarily equal sample sizes, however, the desired transformation may not be readily apparent.

It should be emphasized that the present considerations apply to the more traditional ANOVA approach as well. The two-group problem just discussed provides us with an example. If one wished to form a contrast to test the mean

difference between the groups, the likely choice would be:

$$\hat{\psi}_1 = \bar{Y}_1 - \bar{Y}_2$$

Yet, the results of testing:

$$\hat{\psi}_2 = \frac{1}{2}\bar{Y}_1 - \frac{1}{2}\bar{Y}_2$$

would be identical. Usually, $\hat{\psi}_1$ is the contrast of choice because it is directly interpretable as an estimate of the mean difference. We believe that similar considerations should prevail in the regression approach to ANOVA.

Cohen and Cohen (1975) and Kerlinger and Pedhazur (1973) adopt an approach opposite to the one proposed here. In particular, the code value choices are determined apparently without any consideration of the direct interpretation problem. Should parameter estimates be desired, appropriate (and often complex) transformations of the resulting B_0, B_1, B_2, ... , B_{K-1} are offered. In contrast, with the present approach the researcher decides in advance upon the appropriate coding scheme and a set of code value choices that will permit directly interpretable parameter estimates.

For example, reconsider the two-group problem in Figure 1. Suppose that a researcher wished the resulting B_0 to represent $\bar{Y}_1$ and B_1 to represent $\bar{Y}_2 - \bar{Y}_1$ (code choice a). In terms of the usual regression equation:

$$\bar{Y}_k = \bar{Y}_1 + (\bar{Y}_2 - \bar{Y}_1)X_k$$

where X_k would be the coded value to be applied to the k-th group. It is easy to see that $X_1 = 0$ and $X_2 = 1$ satisfy the respective equalities. Thus, the desired parameter estimates dictate the code value choices.

Alternatively, suppose that one wished B_0 to represent $\frac{1}{2}(\bar{Y}_1 + \bar{Y}_2)$ and B_1 to represent $\bar{Y}_2 - \bar{Y}_1$ (code choice c). Here:

$$\bar{Y}_k = \frac{\bar{Y}_1 + \bar{Y}_2}{2} + (\bar{Y}_2 - \bar{Y}_1)X_k$$

With $X_1 = -\frac{1}{2}$ and $X_2 = +\frac{1}{2}$, the two equalities are satisfied. Once again, the desired parameter estimates dictate the code value choices.

We now extend these notions, with accompanying directions for code value choices, to the K-group ANOVA problem, and then to the analysis of covariance (ANCOVA) and aptitude-by-treatment interaction (ATI) situation as well.

General Procedure for Defining Directly Interpretable Code Values

We now illustrate the general code value determination procedure, as applied to directly intepretable values of B_0, B_1, B_2, ... , B_{K-1} in the K-group one-way ANOVA situation. As indicated above, once the user specifies the effects to be estimated by the K regression coefficients, appropriate code values can be easily determined. The procedure will be illustrated using a particular set of contrasts for K = 3 groups, and then steps will be outlined in a more general sense.

Since one has K - 1 between-group degrees of freedom with which to work, the researcher must first specify a value for the intercept B_0 and K - 1 contrasts for the slopes. A reasonable B_0 choice turns out to be either the grand mean or the mean of one of the groups. Moreover, with the procedure described here, there is no restriction that the contrasts be orthogonal. As a result, this presentation has much more generality than one of the methods of interpretation offered by Cohen and Cohen (1975), which works only for orthogonal comparisons.

Consider a three-group ANOVA problem for which it is desired that:

$$B_0 = \bar{Y}$$

$$B_1 = \bar{Y}_1 - \bar{Y}_3$$

$$\text{and } B_2 = \bar{Y}_2 - \bar{Y}_3$$

With $B_0 = \bar{Y}$, one can then find for each group two code values, X_{k1} and X_{k2}, that satisfy the regression equation:

$$\bar{Y}_k = B_0 + B_1 X_{k1} + B_2 X_{k2}$$

In the case of equal sample sizes[1], $B_0 = \bar{\bar{Y}} = \sum_{k=1}^{3} \bar{Y}_k / 3$, and substituting the two contrasts for B_0 yields:

$$\bar{Y}_k = \frac{\bar{Y}_1 + \bar{Y}_2 + \bar{Y}_3}{3} + (\bar{Y}_1 - \bar{Y}_3)X_{k1} + (\bar{Y}_2 - \bar{Y}_3)X_{k2}$$

For purposes of solution, it helps to collect terms in $\bar{Y}_1$, $\bar{Y}_2$, and $\bar{Y}_3$ on the right side of the equation, as:

$$\bar{Y}_k = \bar{Y}_1(\tfrac{1}{3} + X_{k1}) + \bar{Y}_2(\tfrac{1}{3} + X_{k2}) + \bar{Y}_3(\tfrac{1}{3} - X_{k1} - X_{k2})$$

We then solve this equation for X_{k1} and X_{k2} for each value of k (k = 1, 2, 3). From the last equation, this amounts to choosing X_{k1} and X_{k2} so that on the right-hand side of the equation the targeted $\bar{Y}_k$ has a coefficient equal to one, and the remaining $\bar{Y}_k$ have coefficients equal to zero. For example, for k = 1, we need:

$$\bar{Y}_1 = \bar{Y}_1(\tfrac{1}{3} + X_{11}) + \bar{Y}_2(\tfrac{1}{3} + X_{12}) + \bar{Y}_3(\tfrac{1}{3} - X_{11} - X_{12})$$

Setting the $\bar{Y}_1$ coefficient equal to one yields:

$$\frac{1}{3} + X_{11} = 1$$

$$X_{11} = \frac{2}{3}$$

Similarly, setting the $\bar{Y}_2$ coefficient equal to zero yields:

$$\frac{1}{3} + X_{12} = 0$$

$$X_{12} = -\frac{1}{3}$$

We see, then, that the Group 1 codes for the two independent variables are 2/3 and -1/3, respectively.

Similarly, for k = 2 one can solve the equation:

$$\bar{Y}_2 = \bar{Y}_1(\tfrac{1}{3} + X_{21}) + \bar{Y}_2(\tfrac{1}{3} + X_{22}) + \bar{Y}_3(\tfrac{1}{3} - X_{21} - X_{22})$$

This time, the coefficient of $\bar{Y}_2$ on the right must equal unity, while the other two coefficients must vanish. So:

$$\tfrac{1}{3} + X_{21} = 0$$

$$X_{21} = -\tfrac{1}{3}$$

and:

$$\tfrac{1}{3} + X_{22} = 1$$

$$X_{22} = \tfrac{2}{3}$$

Here, then, the code values for Group 2 on the two independent variables are -1/3 and 2/3, respectively.

Finally, for k = 3 we have:

$$\bar{Y}_3 = \bar{Y}_1(\tfrac{1}{3} + X_{31}) + \bar{Y}_2(\tfrac{1}{3} + X_{32}) + \bar{Y}_3(\tfrac{1}{3} - X_{31} - X_{32})$$

Following the above procedure:

$$\tfrac{1}{3} + X_{31} = 0$$

$$X_{31} = -\tfrac{1}{3}$$

and:

$$\tfrac{1}{3} + X_{32} = 0$$

$$X_{32} = -\tfrac{1}{3}$$

The entire set of code values is therefore given by:

	X_1	X_2
Group 1	$\frac{2}{3}$	$-\frac{1}{3}$
Group 2	$-\frac{1}{3}$	$\frac{2}{3}$
Group 3	$-\frac{1}{3}$	$-\frac{1}{3}$

The procedure described here will work for any set of K - 1 linearly independent between-group comparisons, including the case of K - 1 effect estimates (which themselves can be written as linearly independent contrasts). As mentioned above, such contrasts are to be specified a priori by the user, along with the choice of B_0 (generally either an overall Y mean or a given $\bar{Y}_k$). Direct interpretations of the B_k will then follow from the present procedure. Suppose, for instance, that in the just presented example, one wanted B_0 to represent $\bar{Y}_3$ rather than $\bar{Y}$, with B_1 and B_2 to be interpreted as before. Employing the present procedure yields:

	X_1	X_2
Group 1	1	0
Group 2	0	1
Group 3	0	0

which corresponds to Cohen and Cohen's (1975) dummy code system.

We now summarize the steps of our directly interpretable estimate coding procedure, for the general one-way ANOVA case based on K groups:

Step 1. Specify the desired interpretation of B_0 ($\bar{Y}$ or $\bar{Y}_k$) and K - 1 linearly independent contrast estimators, $\hat{\psi}_i$ ($1 \leq i \leq K - 1$).

Step 2. Write the regression equation in terms of B_0 and the $\hat{\psi}_i$ yielding:

$$\bar{Y}_k = B_0 + \hat{\psi}_1 X_{k1} + \hat{\psi}_2 X_{k2} + \ldots + \hat{\psi}_{K-1} X_{k,K-1}$$

where the X_{ki} are the coded values to be derived for the i-th independent variable in the k-th group.

Step 3. Combine terms in the $\bar{Y}_k$ on the right-hand side of the equation.

Step 4. Solve for the X_{ki} for each value of k to obtain the code values for each group.

Let us use this procedure to define a different set of contrasts for the present three-group problem. Specifically, suppose we wish that: $B_0 = \frac{\bar{Y}_1 + \bar{Y}_2 + \bar{Y}_3}{3}$, $B_1 = \bar{Y}_1 - \bar{Y}_2$, and $B_2 = \frac{\bar{Y}_1 + \bar{Y}_2}{3} - \bar{Y}_3$. Now:

$$\bar{Y}_k = \frac{\bar{Y}_1 + \bar{Y}_2 + \bar{Y}_3}{3} + (\bar{Y}_1 - \bar{Y}_2)X_{k1} + (\frac{\bar{Y}_1 + \bar{Y}_2}{2} - \bar{Y}_3)X_{k2}$$

$$= \bar{Y}_1(\tfrac{1}{3} + X_{k1} + \tfrac{1}{2}X_{k2}) + \bar{Y}_2(\tfrac{1}{3} - X_{k1} + \tfrac{1}{2}X_{k2}) + \bar{Y}_3(\tfrac{1}{3} - X_{k2})$$

For k = 1 we have:

$$\bar{Y}_1 = \bar{Y}_1(\tfrac{1}{3} + X_{11} + \tfrac{1}{2}X_{12}) + \bar{Y}_2(\tfrac{1}{3} - X_{11} + \tfrac{1}{2}X_{12}) + \bar{Y}_3(\tfrac{1}{3} - X_{12})$$

which, since $\bar{Y}_3$ must vanish, yields $X_{12} = +\frac{1}{3}$. Incorporating this into the $\bar{Y}_1$ coefficient results in $X_{11} = +\frac{1}{2}$. Similarly, for k = 2:

$$\bar{Y}_2 = \bar{Y}_1(\tfrac{1}{3} + X_{21} + \tfrac{1}{2}X_{22}) + \bar{Y}_2(\tfrac{1}{3} - X_{21} + \tfrac{1}{2}X_{22}) + \bar{Y}_3(\tfrac{1}{3} - X_{22})$$

Here, $X_{22} = +\frac{1}{3}$ and $X_{21} = -\frac{1}{2}$ satisfy the needed conditions. Finally, for k = 3:

$$\bar{Y}_3 = \bar{Y}_1(\tfrac{1}{3} + X_{31} + \tfrac{1}{2}X_{32}) + \bar{Y}_2(\tfrac{1}{3} - X_{31} + \tfrac{1}{2}X_{32}) + \bar{Y}_3(\tfrac{1}{3} - X_{32})$$

with $X_{32} = -\frac{2}{3}$ and $X_{31} = 0$. The code values can be summarized by:

	X_1	X_2
Group 1	$\frac{1}{2}$	$\frac{1}{3}$
Group 2	$-\frac{1}{2}$	$\frac{1}{3}$
Group 3	0	$-\frac{2}{3}$

Factorial ANOVA designs. The above steps, and resulting interpretations, extend directly to factorial ANOVA designs. Simply define estimates of interest in terms of linear combinations of the C cell means, write the regression

model in terms of B_0 and C - 1 linearly independent between-cell contrasts, and solve for the X_{ci} $(1 \leq i \leq C - 1)$. Because the extension is so straightforward, there is no need to consider it explicitly here.

ANCOVA designs. The ANCOVA model differs from the ANOVA model already described only in that one or more quantitative covariates are added to the regression equation. In the case of only one covariate, one can define a particular group mean as:

$$\bar{Y}_k = B_{C(W)}\bar{C}_k + B_0 + B_1X_1 + B_2X_2 + \ldots + B_{K-1}X_{K-1}$$

where $\bar{C}_k$ represents the covariate mean in the k-th group and $B_{C(W)}$ is the pooled within-group regression coefficient associated with the covariate.[2]

We see that our regression equation cannot balance unless we replace $B_0 + B_1X_1 + B_2X_2 + \ldots + B_{K-1}X_{K-1}$ with $\bar{Y}_k - B_{C(W)}\bar{C}_k$, for then we would have:

$$\bar{Y}_k = B_{C(W)}\bar{C}_k + (\bar{Y}_k - B_{C(W)}\bar{C}_k)$$

If we identify the parenthetical expression as an adjusted mean, $\bar{Y}^A_k$, then from above we have:

$$\bar{Y}^A_k = B_0 + B_1X_1 + B_2X_2 + \ldots + B_{K-1}X_{K-1}$$

which is identical to the previously described ANOVA model, defined in terms of the unadjusted $\bar{Y}_k$.[3] Thus, the methods used previously for determining the coded values will here allow the regression coefficients to estimate adjusted means and contrasts involving adjusted means.

Consider, for example, the previous three-group problem and the dummy coding system described there. In that case, B_0 was to estimate $\bar{Y}_3$, while B_1 and B_2 estimated $\bar{Y}_1 - \bar{Y}_3$ and $\bar{Y}_2 - \bar{Y}_3$, respectively. According to the present ANCOVA coding system, the analogous estimators are given by $B_0 = \bar{Y}^A_3$, $B_1 = \bar{Y}^A_1 - \bar{Y}^A_3$,

and $B_2 = \bar{Y}^A_2 - \bar{Y}^A_3$. Similar parallels hold for other contrast choices.

Designs involving covariate-by-group interactions. The most common educational research design involving covariate-by-group interactions occurs in ATI studies. In such studies, it would be convenient if the estimated parameters could be directly interpreted as differences in slope values between groups. As will now be seen, this can be easily accomplished using the present methods.

Let us examine the model equation for the case of K groups and one covariate. Included in the model are the constant, the covariate, K - 1 coded between-group variables, and another K - 1 coded covariate-by-group interaction variables. Thus we have:

$$\bar{Y}_k = B_{C(k)}\bar{C}_k + B_0 + B_1X_1 + B_2X_2 + \ldots + B_{2K-2}X_{2K-2}$$

Note that in contrast to the ANCOVA model, the covariate coefficient is written as $B_{C(k)}$ rather than as $B_{C(W)}$. This is necessitated by the inclusion of interaction terms in the ATI model. By way of analogy, consider a less than full-rank ANOVA model that includes only main effects. In such a design, only marginal means may be estimated. When the model is of full rank, individual cell means and, thus, interaction effects, may also be estimated. In the present context, the regression model without interaction (ANCOVA) assumes parallelism of the K regression lines and, thus, allows only the pooled within-group regression coefficient for the covariate, $B_{C(W)}$, to be included. When the interaction terms are added, the individual slopes, $B_{C(k)}$, are involved.

With these interaction terms, the equation reduces to:

$$\bar{Y}_k = B_{C(k)}\bar{C}_k + (\bar{Y}_k - B_{C(k)}\bar{C}_k)$$

This equation parallels that of ANCOVA, with the substitution of $B_{C(k)}$ for $B_{C(W)}$.

As before, we identify terms involving B_1, B_2, ... , B_{K-1} with the parenthetical expression, so these coefficients estimate contrasts involving the adjusted means. In this case, however, each group mean is adjusted according to its own covariate, $B_{C(k)}$, thereby rendering the resulting contrasts uninterpretable (since adjusted mean comparisons are interpretable only if the constituent regression lines are parallel). Operationally this means that if one is interested in the K - 1 between-group adjusted main effect contrasts, then two analyses are required: one, based on just the covariate and the first K - 1 coded variables to estimate adjusted mean differences appropriately (as described for ANCOVA); and a second, based on both sets of K - 1 coded variables, with the second set permitting desired contrasts among the K slopes (as will now be described).

In the full model, one needs to choose the interaction codes associated with the slope contrasts so that:

$$B_{C(k)}\bar{C}_k = B_C\bar{C}_k + B_KX_K + B_{K+1}X_{K+1} + \ldots + B_{2K-2}X_{2K-2}$$

where the B_k represent desired contrasts involving the slopes and B_C is a convenient, though not necessarily meaningfully interpreted, choice for the "average" slope. For example, using the same procedure as was applied to the means of our three-group problem, let B_C represent the arithmetic mean of the K = 3 slopes:

$$B_C = \frac{B_{C(1)} + B_{C(2)} + B_{C(3)}}{3}$$

where $B_{C(k)}$ is the regression coefficient relating Y and C within the k-th group. Then, let us say that we want B_3 to estimate $B_{C(1)} - B_{C(3)}$ and B_4 to estimate $B_{C(2)} - B_{C(3)}$. We would have to find code values for X_3 and X_4 for each group to satisfy the relationship:

$$B_{C(k)}\bar{C}_k = [\frac{B_{C(1)} + B_{C(2)} + B_{C(3)}}{3}]\bar{C}_k + (B_{C(1)} - B_{C(3)})X_3 + (B_{C(2)} - B_{C(3)})X_4$$

Collecting terms on the right, we then have:

$$B_{C(k)}\bar{C}_k = B_{C(1)}[\frac{\bar{C}_k}{3} + X_3] + B_{C(2)}[\frac{\bar{C}_k}{3} + X_4] + B_{C(3)}[\frac{\bar{C}_k}{3} - X_3 - X_4] ,$$

an equation which can be solved one group at a time. For the first group, for example, we have:

$$B_{C(1)}\bar{C}_1 = B_{C(1)}[\frac{\bar{C}_1}{3} + X_3] + B_{C(2)}[\frac{\bar{C}_1}{3} + X_4] + B_{C(3)}[\frac{\bar{C}_1}{3} - X_3 - X_4]$$

Since the coefficient for $B_{C(2)}$ must vanish:

$$X_4 = -\frac{1}{3}\bar{C}_1$$

Similarly, since the coefficient of $B_{C(1)}$ on the right must equal $\bar{C}_1$, we need:

$$X_3 = \frac{2}{3}\bar{C}_1$$

Rather than solve for the remaining codes, certain insights may be of use here. First, except for the factors of $\bar{C}_k$, the problem to be solved is identical to that solved earlier in terms of ANOVA contrasts involving means. Second, instead of having to determine the covariate means for each group ($\bar{C}_k$) before setting up X_3 and X_4 (as seems to be required), it is simpler to give each subject in a particular group his or her own (weighted) covariate score. For the present example, the i-th subject in Group 1 would be given a score of 2/3 C_i for X_3 and - 1/3 C_i for X_4, where C_i is that subject's covariate score. When summed across subjects in the design, these codes will produce the desired results. Finally, "mixing and matching" coding schemes for the main effects and interactions is clearly possible under this procedure. As Cohen (1978) has

shown, as long as the constituent main effect has been partialled out, any set of linearly independent codes expressing the interactions may be defined. Thus, one could define dummy codes for the main effect and something other than dummy codes for the interaction, for example. Of course, the most usual choice is to let the interaction codes parallel the main effect codes by taking the product of C_i and each of the first K - 1 coded variables. The point of the present comment is that the "usual" choice is not a necessary one.

Indeed, as may be seen from the above interpretation of interaction regression coefficients, each is a contrast among the $B_{C(k)}$. We have already shown above an example in which B_C was specified as an "average" slope. Unfortunately, this average represents a statistically uninterpretable quantity, being neither the total nor the pooled within-group regression coefficient. Similarly, the only simple contrasts among the $B_{C(k)}$ that are interpretable are pairwise contrasts. Coded variables that produce interpretable combined-group contrasts among means, such as $\hat{\psi} = \frac{1}{2}(\bar{Y}_1 + \bar{Y}_2) - \bar{Y}_3$, would yield uninterptetable contrasts among within-group regression weights.[4] Thus, in such cases, the "usual" procedure for generating interaction-coded variables will often provide uniterpretable estimates and tests. Perhaps the optimal choice for interaction codes would derive from the dummy coding scheme, which would produce a completely interpretable set of estimates and tests. Fortunately, as shown here, this does not restrict one to dummy codes for the mean contrasts, which can be chosen completely independently of the slope contrasts.

Conclusion

To summarize, it is believed that the present procedures will greatly simplify the contrast/effect interpretation problem that presently faces the researcher applying multiple regression coding schemes. That this interpre-

tation problem is a real one has been indicated by our colleagues who have attempted to decode the various coding schemes that are available. For as was pointed out at the outset, although statistical tests of desired effects follow directly from the pages of available texts, interpretations of the properly stated effects do not.

References

Cohen, J. Multiple regression as a general data-analytic system. Psychological Bulletin, 1968, 70, 426-443.

Cohen, J. Partialed products are interactions; partialed powers are curve components. Psychological Bulletin, 1978, 85, 858-866.

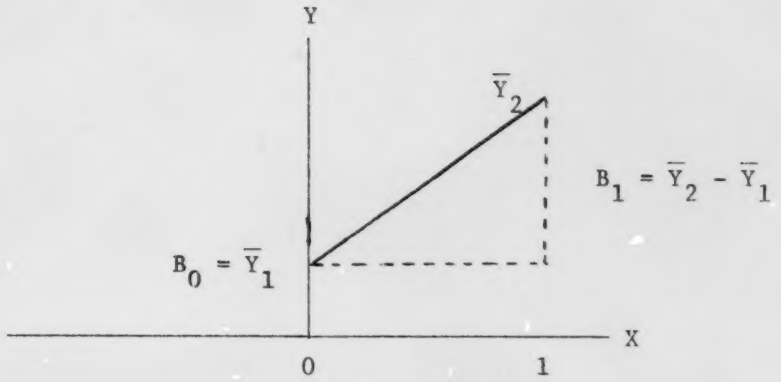
Cohen, J., & Cohen, P. Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, N.J.: Erlbaum, 1975.

Kerlinger, F. N., & Pedhazur, E. J. Multiple regression in behavioral research. New York: Holt & Co., 1973.

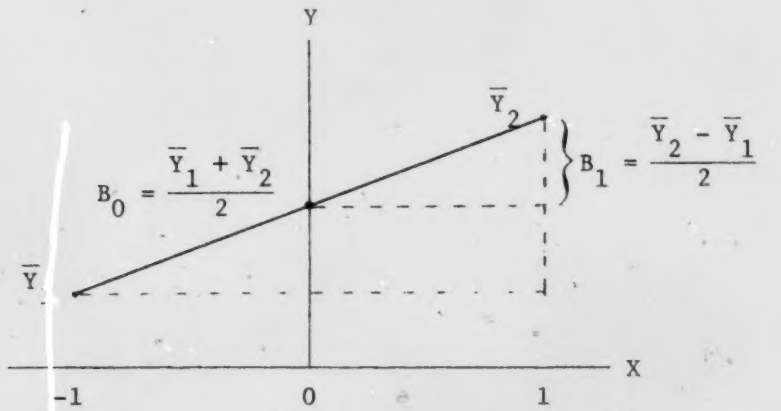
Figure Caption

Figure 1. Example of Different Slope and Intercept Values Resulting From Different Code Choices.

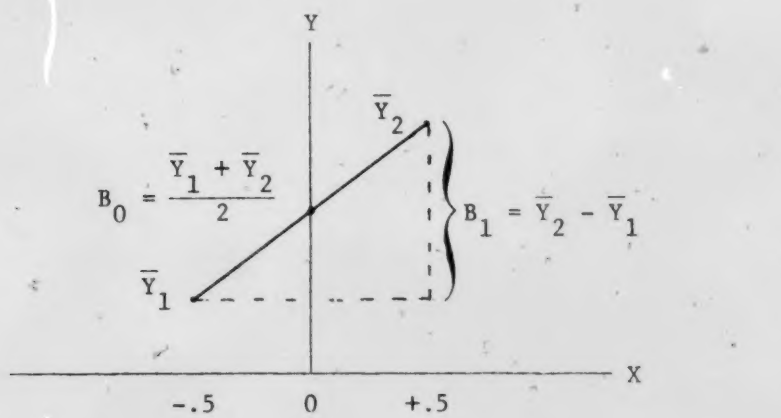
(a)


Y
$\bar{Y}_2$
$B_1 = \bar{Y}_2 - \bar{Y}_1$
$B_0 = \bar{Y}_1$
X
0
1

(b)


Y
$\bar{Y}_2$
$B_1 = \frac{\bar{Y}_2 - \bar{Y}_1}{2}$
$B_0 = \frac{\bar{Y}_1 + \bar{Y}_2}{2}$
$\bar{Y}$
X
-1
0
1

(c)


Y
$\bar{Y}_2$
$B_0 = \frac{\bar{Y}_1 + \bar{Y}_2}{2}$
$B_1 = \bar{Y}_2 - \bar{Y}_1$
$\bar{Y}_1$
X
-.5
0
+.5

Footnotes

The authors are grateful to Beth Brown for typing the manuscript. Requests for reprints should be sent to Ronald C. Serlin, Department of Educational Psychology, University of Wisconsin-Madison, Madison, WI 53706.

[1]For unequal sample sizes, the user must decide upon the particular way in which $\bar{Y}$ and contrasts based on combining the $\bar{Y}_k$ are to be defined (i.e., either as unweighted or weighted means). Once this is done, however, the following procedure still applies.

[2]Because each coefficient in the above equation represents a partial regression coefficient, and because the coded qualtitative variables, $X_1, X_2, \ldots, X_{K-1}$, account for the between-groups effect, $B_{C(W)}$ represents the relationship between Y and C adjusted for the between-groups effect. That is, $B_{C(W)}$ estimates the pooled within-group regression coefficient. Similarly, the B_k estimate the contrasts specified using the above methods, but now adjusted for the covariate.

[3]Note that adjusted individual means in ANCOVA designs are usually defined as $\bar{Y}_k^A = \bar{Y}_k - B_{C(W)}\bar{C}_k + B_{C(W)}\bar{C}$ where $\bar{C}$ is the covariate grand mean. It is of little consequence to include the constant $B_{C(W)}\bar{C}$ term here, however, inasmuch as the contrasts defined will not be affected by its exclusion.

[4]One could, however, use the methods here to estimate and test meaningful combined-group slope contrasts. To do so, one would need to calculate the covariate sum of squares for the k-th group, $SS_{C(k)}$, and then specify the desired slope contrasts by pooling. For example, to compare the pooled slope from Groups 1 and 2 to the slope in Group 3, one would define:

$$\hat{\psi} = \frac{SS_{C(1)}B_{C(1)} + SS_{C(2)}B_{C(2)}}{SS_{C(1)} + SS_{C(2)}} - B_{C(3)} .$$